NFS/RDMA Performance Experimental

Shirley Ma
Oracle

#OFADevWorkshop

Set up, Configuration, Workload

System:
- ConnectX-3, PCIe-Gen3
- X86 Single NUMA node: 6 cores

BIOS configuration:
- Disable Turbo
- Power mode: Performance

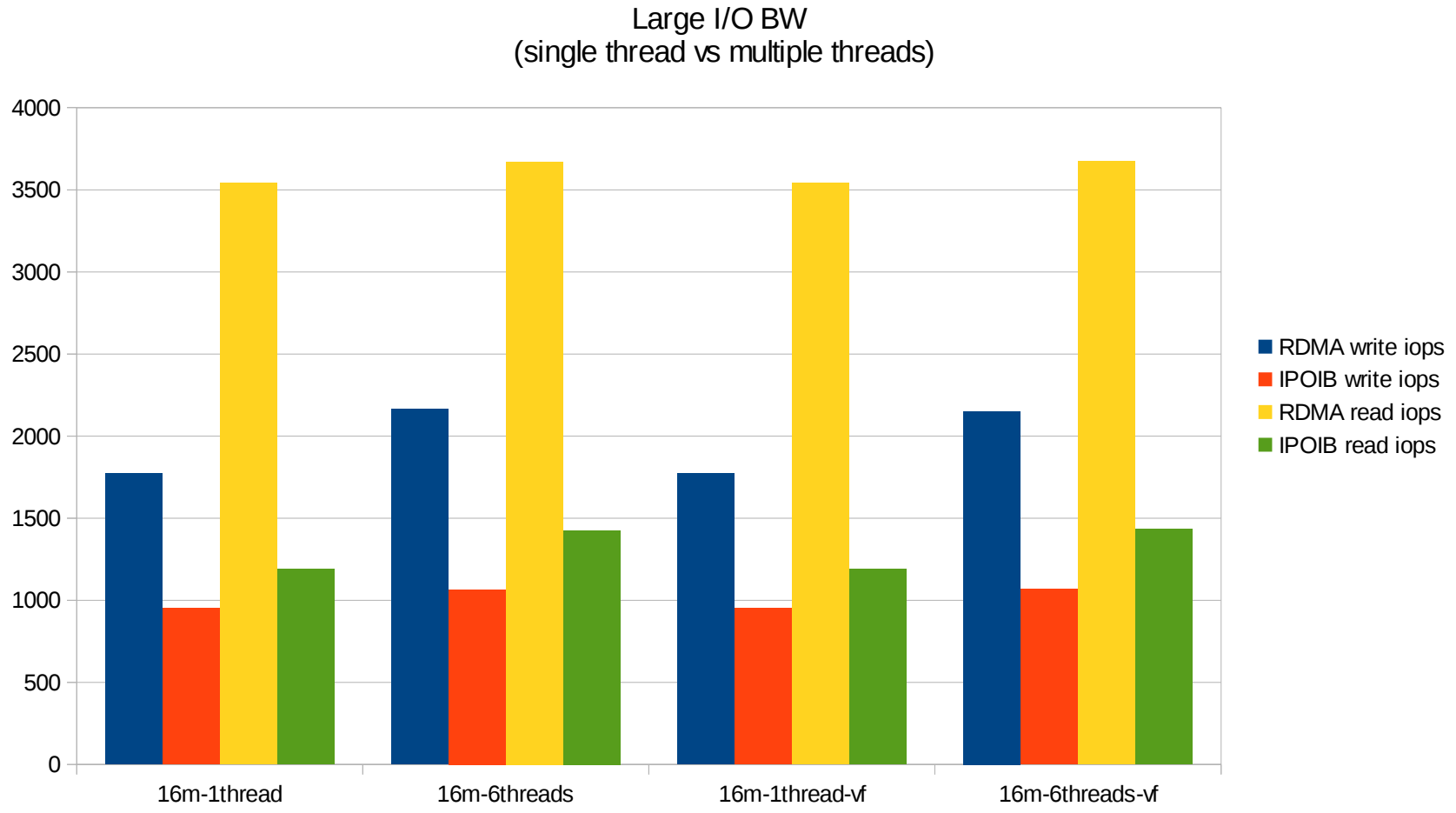
Upstream Kernel:
- Linux 4.0.0-rc1: Client & Server

Benchmark tool: iozone
- Include close in write timing
- Include fsync in write timing
- O_DIRECT
- File size 1G-10G

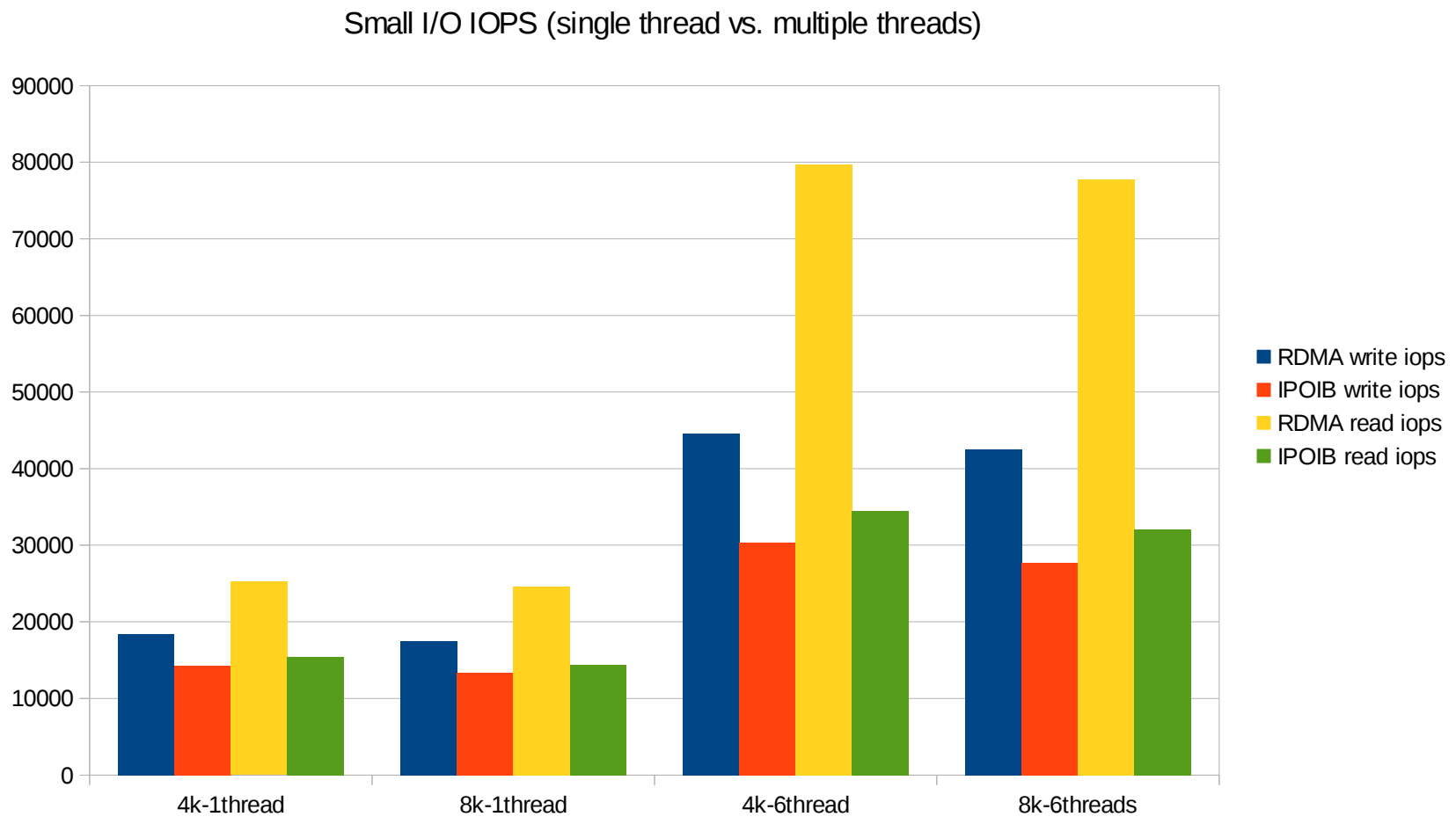
Workload Driven:
- iozone large I/O sequential read/write Bandwidth
- small I/O IOPS read/write (4K, 8K)

NFS Mount option:
- Maximum transport wsize/rsize: nfs/rdma 256K, and nfs/ipoib-cm 1MB
- IPoIB-CM 64K MTU
- NFS V3

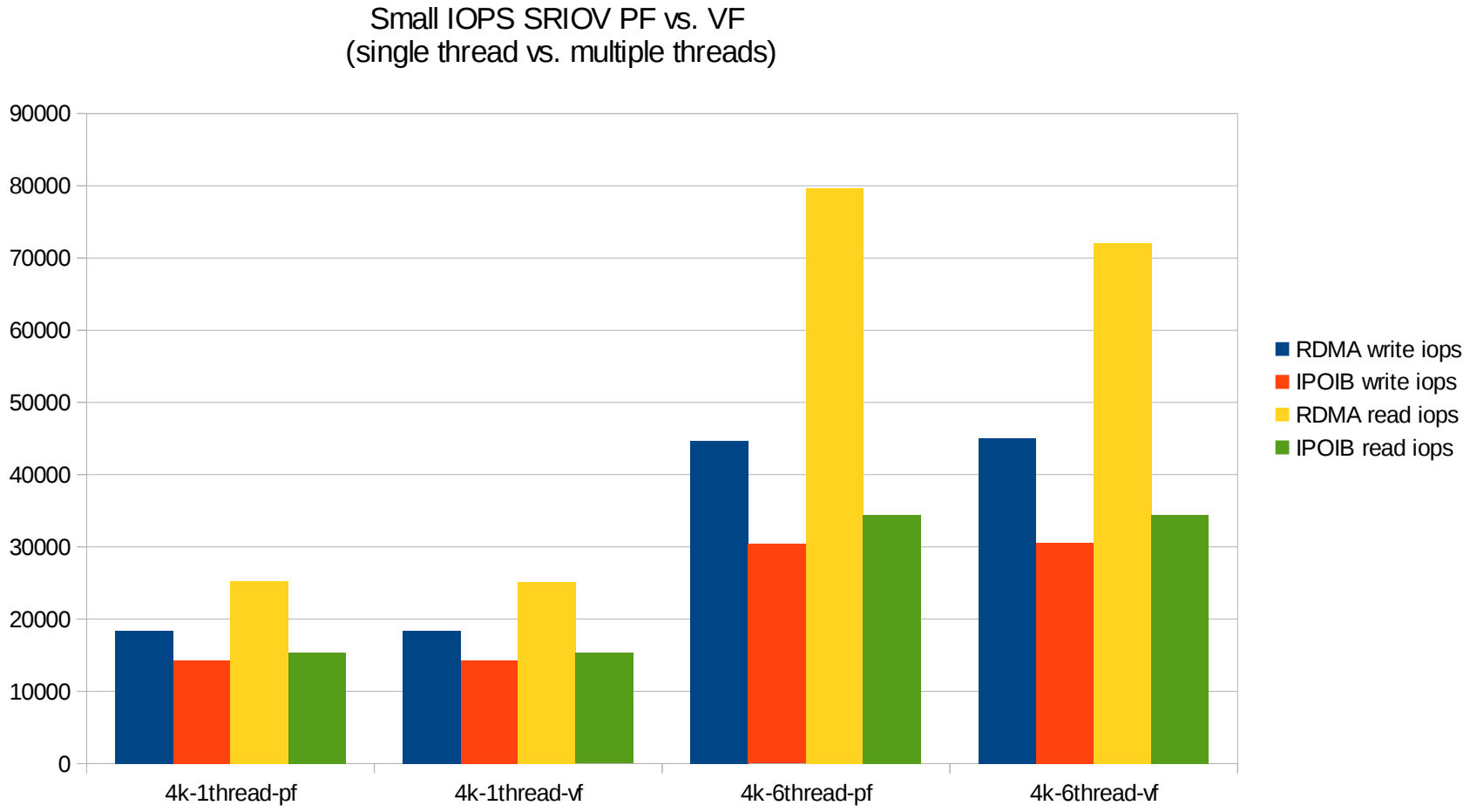
Large I/O Bandwidth (iozone 16M)



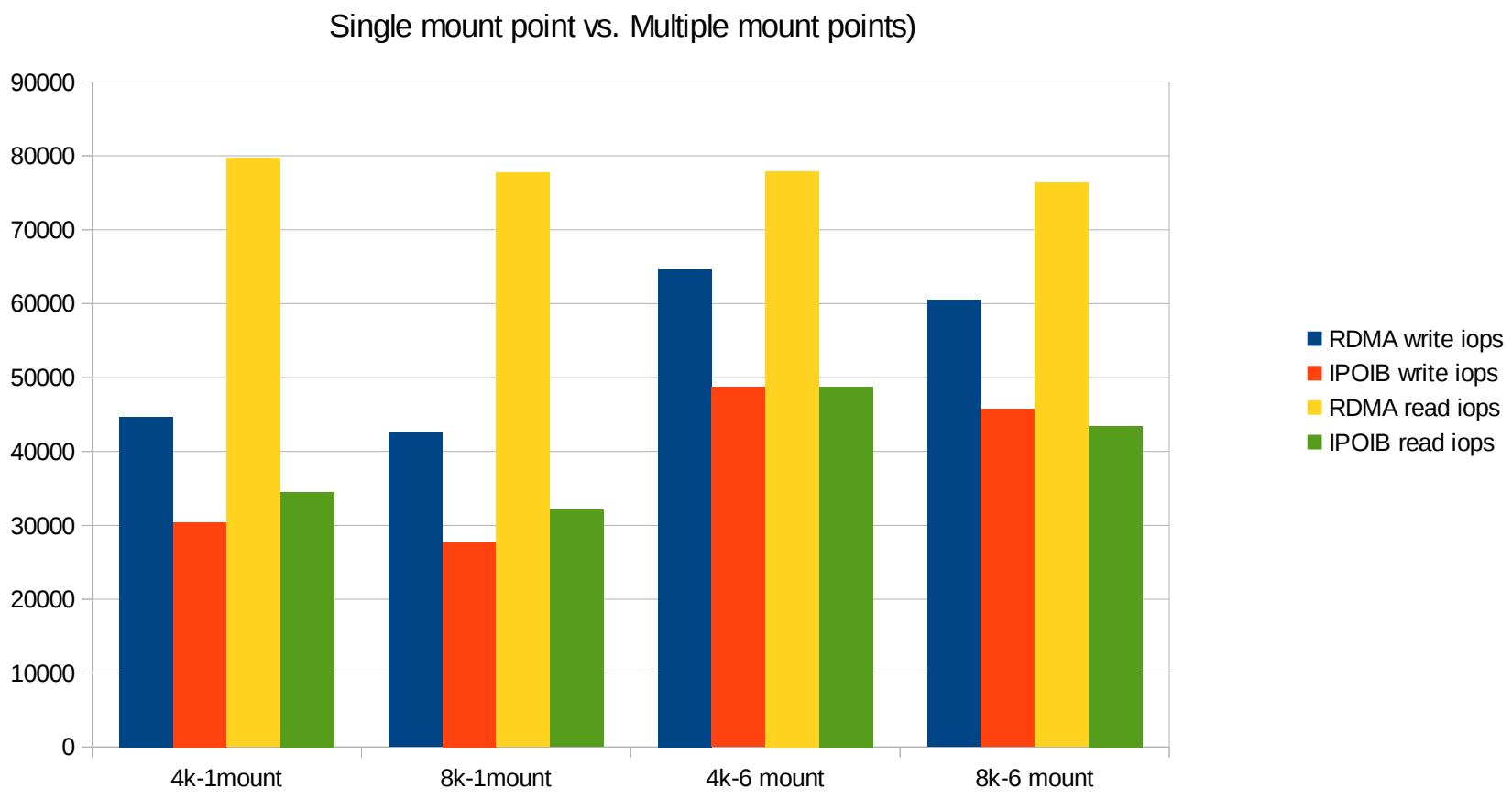
Small I/O IOPS (iozone 4K & 8K)



SRIOV PF vs. VF Small I/O



Single vs. Multiple Connections



Observations

- NFS RDMA outperforms IPoIB-CM (mtu=65520)
- NFS READ (RDMA WRITE) outperforms NFS WRITE (RDMA READ)
- Large I/O outperforms small I/O
- Single CPU is heavily loaded (the one receives the interrupts, /proc/interrupts)
- Small I/O interrupt rates are significantly higher than large I/O

Performance Improvement Experimental

NFS RDMA transport layer – More resources (xprtrdma, svcrdma)

- Increase RPC credit limit from 32 -> 64
- Increase maximum transport wsize/rsize from 256K to 1MB or beyond
- Identify/Reduce resource contention: lock (server side)
- Multiple QPs per mount

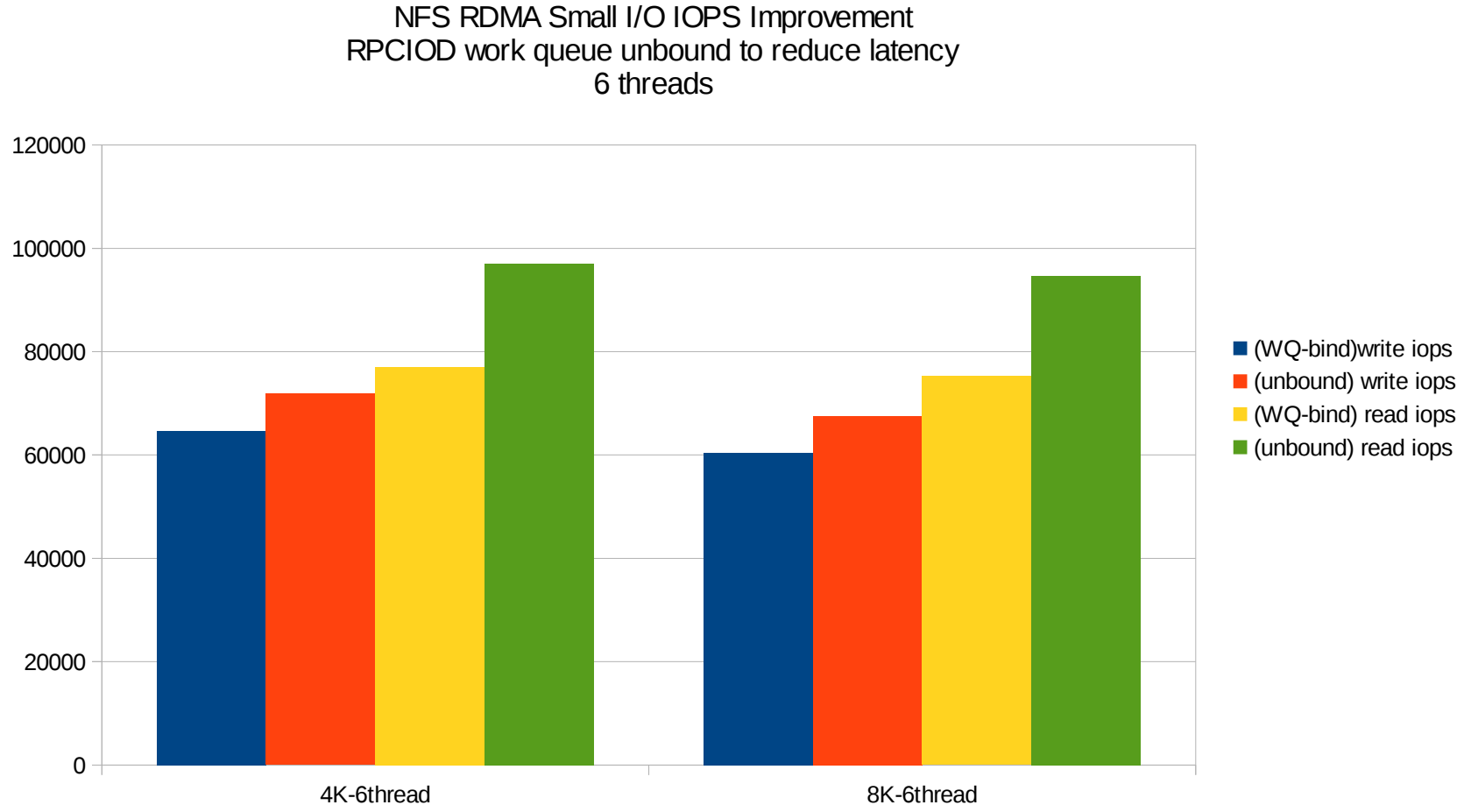
Device interrupts ratio – CPU efficiency:

- Reduce small I/O interrupts rate: increase IOPS/interrupt ratio
- Different completion vector for different QP.
  - irq smp_affinity
  - per CPU CQ

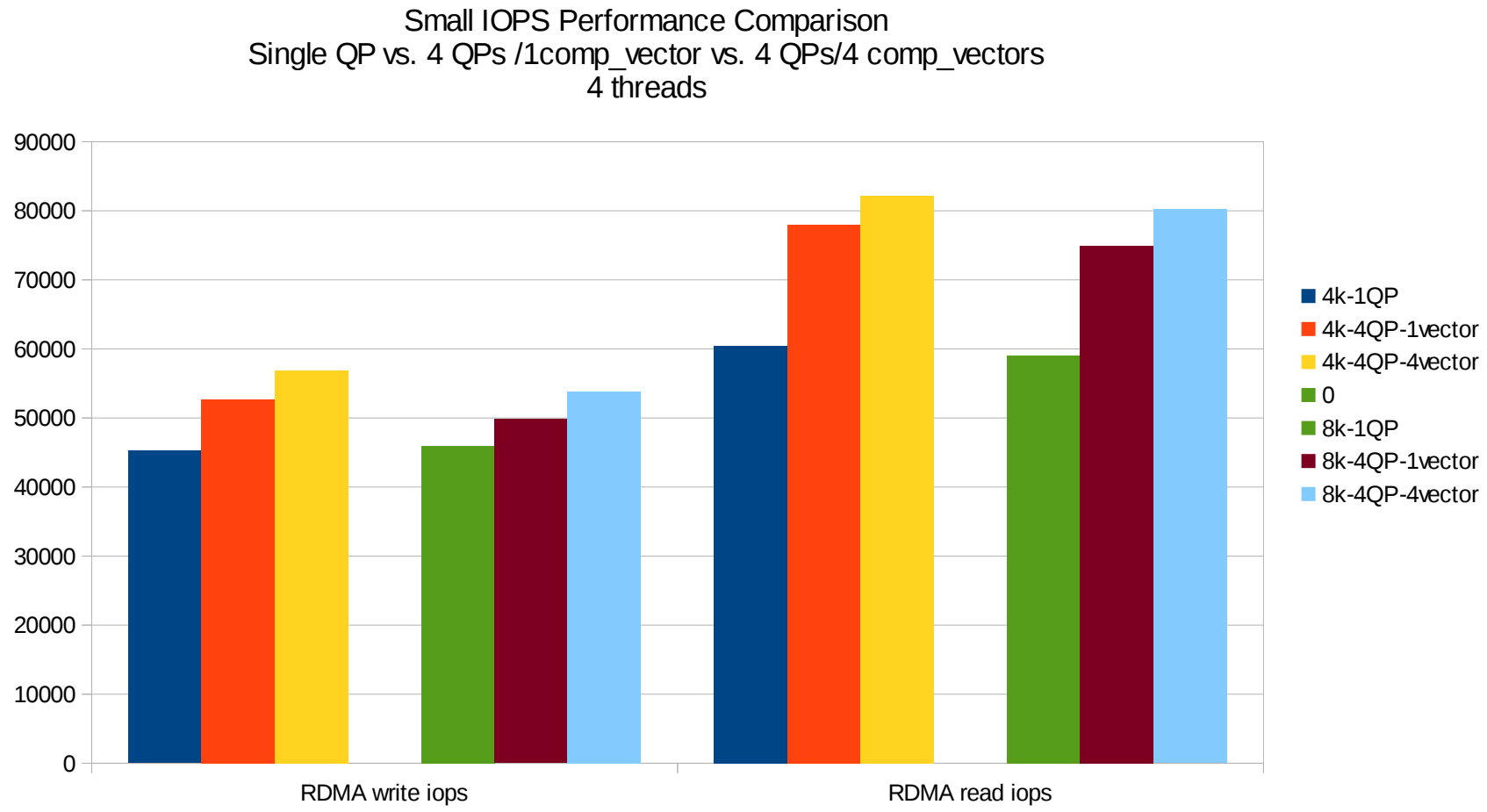
NFS upper layer – Reduce top-down, bottom-up latency:

- Process RPC reply faster: reduce latency
  - Change NFS work queue from bind to unbound (rpciod, svcrdma, nfsd)
  - Reduce scheduling latency when there are other CPU intensive workload
- Reduce resource contention: spin lock, single thread utility bottlenecks
  - Identify the contention

Reduce latency from NFS layer
(RPCIOD Work Queue Bind -> UNBOUND)



Small I/O IOPS 1QP vs. 4QP
(4 threads NFS RDMA read/write)



Thank You

OPENFABRICS ALLIANCE

11th ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP